AO05: Satellite-Based Cloud-Phase Detection over Greenland with Machine-Learning

Name: Alex Dobra. **Supervisors**: Dr Rui Song, Professor Don Grainger

Abstract

This study presents a machine learning (ML) based approach to cloud detection and cloud-phase classification over Greenland, a notoriously difficult region for existing passive imager detection and classification (DC) algorithms. Two Random Forest (RF) models, for daytime and nighttime, were trained separately on Moderate Resolution Imaging Spectroradiometer (MODIS) spectral observations in a five-year period (2013–2017), and were validated against Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) data. The RF models achieved significant improvements in performance compared to the currently operational MODIS cloud DC algorithms: 83.83% agreement with CALIOP for the RF nighttime model and 87.98% agreement for the RF daytime model. The success of these ML models in Greenland suggests a promising avenue for enhancing cloud classification in remote sensing, though challenges remain in generalising these models across all permanent ice surfaces, as evidenced by varied performance in Antarctica in the highest altitude regions. More attention is needed in designing physically relevant model inputs that can distinguish thin and low-lying clouds from the surface.

1 Introduction

Clouds are a fundamental component of the Earth's atmosphere and significantly influence radiation balance and hydrological cycles, making their accurate detection and classification important for a multitude of atmospheric science pursuits. This work is meant to aid the effort in creating long-term, high quality cloud property datasets by synergising information from different Earth Observation missions, an effort spearheaded by the European Space Agency's Cloud Climate Change Initiative.

Cloud mask and phase information is often the first step in obtaining cloud properties used in climate predictions — like cloud optical depth, cloud effective radius, and cloud-top temperature (see Baum et al., 2012) — and so, inaccuracies in cloud-phase classifications lead to large uncertainties in predictions. Additionally, misclassification of cloud cover poses challenges to aerosol detection algorithms, as even minimal cloud contamination can lead to inaccurately high aerosol optical depth (AOD) retrievals, significantly impacting downstream retrieval products and scientific analyses, as detailed by Remer et al., 2005.

A particularly notable cloud-phase detection and classification (DC) product, and relevant to this study, is the Moderate Resolution Imaging Spectroradiometer (MODIS) Cloud Properties product. Operational algorithms for cloud-masking (Frey et al., 2008; Ackerman et al., 2008) and cloud-phase classification (Platnick et al., 2017; Marchant et al., 2016) have been developed, but they exhibit significant weaknesses, especially in classifying cloud-phase over permanent snow or ice surfaces. MODIS has been shown to misclassify as much as 20% of clouds over the Arctic and as much as 30% over Greenland as clear (Chan and Comiso, 2013). Statistics from this work corroborate these findings, as shown in figure 1.

		MYD06 IR CALIOP			MYD06 OP CALIOP		
		Clear	Water	Ice	Clear	Water	Ice
MODIS	Clear	0.86	0.17	0.40	0.91	0.08	0.46
	Water	0.01	0.09	0.02	0.04	0.70	0.07
	Ice	0.07	0.09	0.39	0.05	0.11	0.44
	Undet	0.06	0.66	0.20	0.01	0.11	0.04

Figure 1: Contingency table showing MODIS vs CALIOP phase agreement over Greenland, 2013–2017, for the two MODIS algorithms using infrared and optical bands, respectively. The overall phase agreement of the all-day MODIS algorithm with CALIOP is 52%, and that of the day-only algorithm is 71%.

These algorithms include decision-tree structures and voting systems involving manually-tuned tests and thresholds that are empirically selected based on the developer's experience and access to validation datasets. The manually-set thresholds are sensitive to instrument-specific characteristics, such as spectral band-pass, noise profile, and viewing angle. Additionally, the geographical and temporal limitations of data used in the tuning process pose challenges for global or annual applications, with noticeable biases when applied beyond the region or season the algorithm was made for. Even if the data used for tuning is global, the rigidity of the decision-tree means that performance drops in regions with extreme surface

conditions, such as Greenland.

In contrast, machine-learning (ML)-based DC algorithms could prove to be an e cient and exible approach. By nding non-obvious relationships between predictors and the classi ed categories, ML models eliminate the need for manually de ned thresholds or predetermined spectral patterns that have to be matched to speci c atmospheric features. This allows larger and more diverse feature-sets to be take into consideration in the classi cation at almost no incremental cost. This has been successfully achieved in many remote sensing applications, but most notably, and relevant to this work, examples include cloud-phase classi cation | Wang et al., 2020 | and aerosol classi cation | Lee et al., 2021.

This project aims to determine the applicability of ML-based cloud-phase classi cation over Greenland, a region that poses particular di culties to the current hand-tuned DC algorithms. The di culty in accurate cloud-phase detection is caused by the high re ectivity of the high-altitude Greenland ice sheet and the poor contrast between the surface and clouds in the infrared bands. Cloud-Aerosol Lidar with Orthogonal Polarisation (CALIOP) data will be considered ground truth and used for the evaluation of the ML model. Active instruments are widely used for validation in remote sensing, as they often o er a di erent suite of physically relevant observations than passive sensors, and CALIOP in particular is considered to be one of the best space-based cloud detection systems for the Arctic region (Chan and Comiso, 2013). Descriptions of instrument capabilities will be given in section 2, along with the methods used for compiling and cleaning the training and validation data. Details on model training and evaluation are presented in section 3. In section 4, limitations of the models will be discussed, along with further work to be done. The conclusions are presented in section 5.

2 Data

This section will go into detail on how CALIOP measurements are used as reference labels for the ML models, how MODIS data is used as the model input and present the performance of the hand-tuned decision tree algorithm implemented by the MODIS atmosphere team over Greenland. All satellite data used in this work has been taken from the Centre for Environmental Data Analysis (CEDA) archive connected to the JASMIN supercomputer, the UK's data analysis facility for environmental science.

2.1 Collocation

Collocation is the process of nding the measurements from both satellites that are in the same place, at roughly the same time. Because MODIS and CALIOP are on board satellites that share an orbit, passing above the same location with just a two-minute delay, this step is relatively straightforward. For each CALIOP orbit with data over Greenland, MODIS images captured within the relevant time interval are searched for, taking the time delay of the satellites into consideration. Once the MODIS images that might completely envelop the CALIOP measurements are found, each CALIOP measurement is assigned a single MODIS pixel based on the closest distance. See gure 2 for a graphical representation.

Figure 2: Collocation of CALIOP measurements and MODIS pixels over Greenland. The pixel sizes are exaggerated and their density is reduced for illustration purposes. Two MODIS images, shown in di erent colours, were needed for the collocation of the full CALIOP overpass in this case.

Distances are calculated with the formula

$$D = R\sqrt{^2 + (\cos\ _m\ )^2}; \qquad (1)$$

where R is Earth's average radius, and are the latitude and longitude di erences in radians, and $_m$ is the mean latitude. This is a at-surface distance formula that takes into account the variation in distance between meridians with latitude, and it is a fast and fairly accurate approximation for points on the Earth spheroid not far from each other. The combination of granule timestamp (for example 2017-01-01T13-23-09 for CALIOP and A2017001.1325 for MODIS) and measurement ID within the le is saved in a collocation databasefor later use, and it uniquely links every CALIOP measurement over the icy surface of Greenland to the

closest MODIS pixel in both space and time.

2.2 Reference labels of pixels

Active instruments are widely used for validation in remote sensing, as they are considered to be more accurate than passive sensing instruments at identifying atmospheric features. Cloud-phase classi cations from the CALIOP instrument are used as reference labels for the ML models in this study due to its accuracy and long mission time.

2.2.1 CALIOP instrument description

CALIOP is a two-wavelength lidar (532 nm and 1064 nm) providing high-resolution vertical pro les of the atmosphere. The sensor is on board the Cloud-Aerosol Lidar and Infrared Path nder Satellite Observations (CALIPSO) satellite, which was part of the A-Train satellite constellation from 2006 until 2018. The A-Train is a group of satellites with complementary cloud-observing capabilities, following very close orbits to each other. CALIPSO is in a nearly-polar (98.22 degrees inclination), circular sun-synchronous orbit, crossing the equator northwards at about 13:30 local time (Winker et al., 2009). The instrument uses laser pulse time-of- ight to determine the position of atmospheric features, with a vertical resolution of 30 m in the troposphere and 60 m in the stratosphere. The lidar measures attenuated backscatter intensities at both wavelengths and depolarisation ratio for the 532 nm band. This data is used to accurately identify atmospheric features like aerosols and clouds.

2.2.2 Data product description and label quality control

This section will present the relevant data in this product and how a quality control process is applied to produce accurate reference labels for the ML model. Some needed terminology:

- Level 1 refers to information obtained directly from the satellite, with minimal processing (e.g. calibration adjustments).
- Level 2 refers to data products that have been obtained as a result of extensive processing after the data has been beamed back to Earth. The spatial and temporal formats of the data are preserved.
- Pro le refers to all vertical information retrieved by CALIOP at a certain location and time. In level 1 data, a pro le will contain various properties like perpendicular and parallel backscatter co- e cients as a function of height, geolocation, time and spacecraft geometry information. In level 2 data products, a pro le can have multiple layers and additionally hold column properties, such as column aerosol optical depth.
- Layer refers to an atmospheric feature (aerosol or cloud) detected by the level 2 processing algorithms, and contains spatial and optical characteristics of the feature found. Examples of spatial features are layer base and top altitudes. Examples of optical layer properties are integrated attenuated backscatter and optical depth. Furthermore, each layer is classi ed as aerosol or cloud, and subclassi ed into aerosol type (such as smoke, dust, sulphites etc.), cloud type (such as altocumulus, cirrus etc.) and more importantly for this project, cloud-phase. The CALIOP cloud-phase categories are Water, Ice , Oriented ice and Unknown For a simpli ed, albeit ctitious, representation of layers within pro les, see gure 3.

Figure 3: Illustration of CALIOP pro les and layers. The number of cloud and aerosol layers in each pro le is shown. The method of assigning classi cations to each pro le using this information is explained in detail in appendix A.

The CALIOP data used in this study comes from the level 2 5 km \Merged Layer" product, version 4.51 (newest at the time of writing). The CALIOP layer detection algorithm is described in Vaughan et al., 2009, and descriptions of the datasets contained within the LID_L2_05kmMLay-Standard-V4-51 product can be found on the NASA CALIPSO mission website, NASA-LaRC, 2024.

A multi-step quality control process is applied to the CALIOP data to ensure that the reference label accuracy is as high as possible. This process is a stricter and expanded version of the one presented in Wang et al., 2020. At the end of the process, each pro le should be classi ed as one of the Water, Ice , Clear categories (see gure 4 for a diagram). For example, in the situation presented in gure 3, pro les 2, 3, and 7 and contaminated by aerosols, and eliminated. Pro le 6 is Clear . Pro les 1 and 4 are Ice , and if the aerosol layer in pro le 7 turns out to be

of negligible optical depth, pro le 7 would be Water. Pro les 5 and 8 are classi ed asMixed multilayer at rst, and subsequently reclassi ed as Water, Ice or Ambiguous A more detailed breakdown of the process is given in appendix A.

Figure 4: The quality control process applied to collocated measurements of CALIOP and MODIS. The number of remaining datapoints N is shown at every step.

2.3 Model input

Data from the MODIS instrument on board NASA's Aqua satellite orbiting in the A-Train is used as the model input. MODIS has been chosen for this project because it has been producing high-coverage rich spectral information relevant to cloud properties throughout its long mission time (20+ years), and because of its shared orbit with the most accurate space lidar sensor to date, CALIOP.

2.3.1 MODIS instrument description

MODIS is a whisk-broom sensor, in which a rotating mirror is scanning across-track, collecting a small array of along-track pixels at a time. The sensor provides radiance data for each pixel in 36 bands across the optical and infrared spectrum (410 nm to 14µm), at various spatial resolutions. The constant stream of data is divided into granules for ease of data handling. A typical granule at 1 km ground resolution is an array of 2030 by 1354 pixels, but this study is using 5 km MODIS products (or subsampled 1 km products), which results in images with a size of 406 by 270 pixels.

2.3.2 Data product description and MODIS pixel quality control

In nighttime, only MODIS's emissivity bands (20{36, excluding 26) can be used. In daytime, the re ectivity bands (1{19 and 26) can also be used. The radiance data is subsampled at 5 km resolution from the level 1 1 km product MYD021km. Cloud top temperature and pressure are often strong predictors of cloud-phase, and so these are obtained from the cloud-property level 2 MODIS product, MYD06_L2. These values are not to be trusted blindly, as they

depend on the current MODIS cloud-phase classi cation algorithm. However, they are useful additional information the model can take into consideration. Surface temperatures contribute to the retrieved infrared radiances and interfere with the identi cation of cloud-phase. Thus, the surface temperature during daytime and nighttime is obtained from the level 3 8-day composite product MYD11A2, and added as a feature of the dataset. Having knowledge of the surface temperature in conjunction with top-of-atmosphere radiances should give more cloud-surface discrimination power to the model. Geolocation, timing and geometry information is also retrieved. This includes the latitude, longitude, viewing zenith angle (VZA) and solar zenith angle (SZA) for every pixel. VZA has been shown to a ect many of the cloud properties retrieved by MODIS (Maddux et al., 2010) and is an unavoidable artefact, as is SZA dependency, and are important enough to include. Finally, only pixels over permanent snow or ice are kept. The surface type can be obtained from both CALIOP and MODIS datasets, but the annual L3 MODIS product MCD12C1 was chosen because of its reliability compared to the instantaneous CALIOP surface classi cation.

2.4 Existing MODIS cloud-phase classi cation algorithms

The MODIS cloud product MYD06_L2 contains a variety of useful cloud properties, including cloud-phase classi cation for every pixel. There are two main algorithms producing these classi cations. The rst only uses infrared bands and produces classi cations in both daytime and nighttime; this will be referred to as the MYD06 IR algorithm from now on, and it is explained in detail by Baum et al., 2012. The second algorithm additionally makes use of optical and short-wave infrared bands, and is only applied to pixels captured in daytime | de ned as pixels with a SZA value of less than 81.36 degrees; this will be referred to as the MYD06 OP algorithm from now on, and it is explained in detail by Platnick et al., 2017. The phase agreement of these algorithms with collocated CALIOP classi cations over the ice surface of Greenland during a ve year period (2013{2017) is shown in gure 1. The MYD06 IR algorithm classi es a large fraction of CALIOP{identi ed water clouds as Undetermined, and both algorithms have trouble distinguishing clear scenes from ice clouds. The overall phase agreement of MYD06 IR with CALIOP is 52%, and that of the MYD06 OP algorithm is 71%, suggesting that the extra information found in the optical bands signi cantly improves the classi cation ability. Apart from the inherent limitation of the rigid structure decision tree algorithms that do not take into account extreme surface conditions, the poor discrimination between clear and ice-cloudy pixels can, to a large extent, be attributed to a severe defect in the MODIS Aqua 1:6 µm detector. The 1.6 micron band plays a crucial role for cloud-property retrieval over snow and ice surfaces (see Platnick et al., 2001) since snow typically has much lower albedo in this band. E orts have been made, successfully, to restore the MODIS Aqua 1.6 micron band (see Gladkova et al., 2011) using neighbouring band information, but due to time constraints, the technique was not used in this work.

3 Model training and validation

Two models were trained: a nighttime model only using infrared bands, and a daytime model using both infrared and optical bands. This is done both to get the most out of the optical information during the day and to have the ability of direct comparison to the MYD06 IR and MYD06 OP algorithms. Subsection 3.1 introduces the ML model chosen for this project, 3.2 details the train-test data splitting procedure, 3.3 presents the choices made in feature engineering, 3.4 the process of feature subset selection, 3.5 details hyperparameter tuning and 3.6 presents the nal model performances.

3.1 Model selection and parameters

There is a wide range of \o -the-shelf" machine learning models to choose from for classi cation problems, and the selection and parameter tuning is both an art and a science. The choice for this project is a Random Forest Classi er (RF), as inspired by a successful ML-powered approach at cloud-phase classi cation by Wang et al., 2020. Random forests have been shown to largely correct for the pitfalls of single decision tree classi ers (Ji and Ma, 1997).

RF is an ensemble classi cation technique, using multiple instances of a basic classi er unit, the decision tree. A decision tree is composed of nodes and leaves: a leaf contains one of the possible classi cation results, and a node contains a binary condition on a feature of the dataset that subsets the space of datapoints, for example a condition might be BT(11) > 270 K, where BT(11) stands for the brightness temperature of the 11 micron band. When using a decision tree to predict the class of a particular datapoint, starting with the root node, a route is followed that meets the condition at each node, until it eventually reaches a leaf containing the classi cation result.

In the RF algorithm, the trees are grown independently on random samples of the data set, and a set of tricks is used to decorrelate the trees. More details on how the RF algorithm works are given in appendix B, along with a visualisation of the rst few levels of a decision tree.

The RF has several parameters that need to be hand-picked based on the situation. A few parameters (tree depth, number of trees in the forest and minimum number of samples at a leaf node) will be varied for the highest preforming feature set.

3.2 Train-test data split

To make sure the performance of the model is evaluated without bias, the data that has survived quality control is divided into three groups, training, validation and test. The training set is only used for training the models, the validation set is used to evaluate model performance in the intermediary steps (feature selection and hyperparameter optimisation) and the test set is only used at the end, to provide the nal evaluation for the model. The method of splitting needs to be given some thought. A natural suggestion is to use some full years of data for the training and some di erent years for the test (for example, train on 2013-2016, test on 2017). However, instrument electronics characteristics drift as they age and might introduce, over long periods of time, signi cant biases. The opposite approach, picking a certain percentage of measurements at random to be in the test set and letting the rest be in the training set is also problematic, because of spatial autocorrelation. Given the spatial nature of the data, measurements close to each other will have a high degree of correlation (for example multiple measurements in di erent areas of the same cloud). This e ectively \leaks" information from the test set into the training set, as many data points are too similar. This is a known problem in geosciences, see for example Ploton et al., 2020, and causes the overestimation of the predictive performance of the model. To avoid this problem, the split of training and testing data is done as following: for every block of 20 consecutive days, days 19 to 20 (roughly 10%) are put in the test set and not seen until the nal model evaluation. The remaining data is split into training and validation sets in a similar way: for each 20-day block, days 17-20 (roughly 20% of non-test samples) are put in the validation set and the rest are put in the training set. The resulting split is graphically represented in gure 5 for a selected 3-month stretch.

Figure 5: Graphical representation of the training (blue), validation (orange) and test (green) data splits. Each stripe represents a full day of data.

3.3 Feature engineering

The creation of a new set of predictors derived from the original data can sometimes help the model recognise deeper relationships that could only be explored by increasing the number of internal parameters. In the context of this project, there are physically relevant quantities that could prove better predictors than just the radiances from MODIS. Drawing inspiration from the MYD06 algorithms (Baum et al., 2012 and Platnick et al., 2017), several engineered predictors were added to the dataset, and they proved to be quite relevant. These are the brightness temperatures of selected bands (7µm, 8:5µm, 12µm and 11µm) obtained from radiance values by inverting Planck's law, see equation 2.

$$BT(\ ) = \frac{hc}{k_B} \ln^{1}\ 1 + \frac{2hc^2}{R\ ^5}\ ; \qquad (2)$$

where R stands for the radiance at wavelength . The terms h, c and k_B are physical constants with their usual meaning.

The brightness temperature di erences between several bands are also used in the MYD06 IR algorithm, for example BTD(8:5; 11) = BT(8:5) BT(11) (sensitive to ice clouds) and BTD(7:3; 11) = BT(7:3) BT(11) (helps separate high clouds from low clouds). The NASA researchers have identi ed even better predictors, the so called beta parameter (invented by Parol et al., 1991), de ned as $= \frac{\ln(1\ \ "_y)}{\ln(1\ \ "_x)}$, where $"_x$ and $"_y$ are the cloud emissivities in bands x and y. Unfortunately, there are a lot of missing pixels in the MODIS emissivity datasets and so the beta parameter could not be reconstructed in time.

3.4 Feature selection and importance

Some features have more predicting power than others, and some features might add needless complexity and noise to the data. This warrants designing a process that selects the most relevant features. In theory, all possible subsets of features should be tested and the one that gives the highest performance is chosen. In practice, this is an impossibility when dealing with a large number of available features N, as the number of unique subsets is $^N_2$. To reduce the cost of this search, a technique called Stepwise Feature Selection is used. At each step, a pool of unused predictors is considered. From this set, a single predictor producing the largest increase in the performance metric is chosen and added to the model. This is repeated until the desired number of predictors is reached. Appendix C contains gure 11, a diagram of the process.

The best model in a set is chosen by using a performance metric. This cannot be the training accuracy,

as it will lead to over tting on the training set. The test accuracy cannot be used either, because information about the test set is leaked into the model before it is supposed to see it | and the predictors that happen to be the best for the test set will be chosen. A separate validation set is used for this scoring, obtained with the method described in the previous section, to reduce the e ect of spatial autocorrelation. The stepwise feature selection algorithm, while a guided search through model space, is not guaranteed to nd the best model out of 2^N, but is worth the reduction in computational cost.

Ten features were selected for each model type; see gure 6 for the ordered lists of selected predictors, from the most to the least important.

Figure 6: The sets of ten predictors arrived at by the Stepwise Feature Selector for each model, in the order of their relative importance. The importance score of a feature can be roughly interpreted as the loss in model accuracy, were that feature to be removed.

3.5 Hyperparameter grid search

Having found the optimal feature set using the stepwise feature selection algorithm, further optimisation can be achieved by tuning the model parameters. The most straightforward method is a hyperparameter grid search, where the model is trained and scored (on the validation set) for many values of model parameters. The three parameters varied in this search were maximum tree depth N_{depth}, with values from 10 to unlimited, number of trees in the forest N_{trees}, with values from 50 to 500, and the number of minimum samples at a leaf node N_{leaf}, with values from 1 to 20. See appendix D for the model performance for these values. The parameters that gave the highest validation scores were N_{depth} = 40, N_{trees} = 500 and N_{leaf} = 2 for the nighttime model and N_{depth} = Unlimited, N_{trees} = 500 and N_{leaf} = 2 for the daytime model.

3.6 Results

Training the models with the best features and with the optimised hyperparameters and scoring them on the previously unseen test set yields a 85.23% accuracy for the RF nighttime model and 89.06% accuracy for the RF daytime model. See gure 7 for the confusion matrices. For a more realistic performance, an Undetermined category can be added back; pixels classi ed as one of the Water, Ice or Clear categories with low con dence (probability less than 50%) are moved to the Undetermined category. This results in 83.83% accuracy for the RF nighttime model and 87.98% accuracy for the RF daytime model, with about 3% of pixels being moved to the Undetermined category due to low con dence.

Figure 7: The performance of the two models over Greenland.

The performance of the Greenland-trained models was also tested on data over Antarctica in the same time period, 2013 to 2017. The Antarctica data went through the same quality control and feature engineering processes described previously. The RF nighttime model exhibits a signi cant reduction in predictive performance (down to 75.39%) while the RF daytime model scored 85.23%, a more modest decrease. Possible reasons for the performance drop and other limitations of the project are discussed in the next section.

4 Discussion and further work

While the classi cation performed by both RF models signi cantly outperform the MYD06 algorithms, the project has some limitations. The rst is that the model cannot be con dently applied to all MODIS pixels in an image, but only to a narrow band (± 10 degrees) around the nadir pixels. This is because CALIOP is a near-nadir (3 degrees) viewing instrument, and due to its shared orbit with MODIS, most collocated pixels have a small MODIS view zenith angle (VZA). As it has been established by Maddux et al., 2010, MODIS data exhibits VZA dependency, therefore making any classi cations at VZA values outside of those in the training range less certain. This is not as drastic for permanent ice

and snow surfaces, as their location at high latitudes means that there is a relatively large density of orbit overpasses, and even thin strips around the MODIS nadir pixels produce decent coverage | this is not the case for mid and low latitudes.

Investigating the properties of clouds which have been misclassi ed by the RF models should give some clues on the kinds of spectral predictors to be added for increased performance. The RF models struggle to correctly classify all but the highest and thickest clouds | of all misclassi ed clouds, most are thin and close to the surface. The high-altitude thin clouds are also frequently missed, especially by the daytime model. Figure 8 shows smoothed histograms of the correctly classi ed and misclassi ed pixels over Greenland.

Figure 8: Two-dimensional kernel density plots of the correctly and incorrectly classi ed cloudy pixels.

The experiment over Antarctica has also shown that the Greenland-trained models, while still outperforming the MYD06 algorithms, are not as generally competent at classifying cloud-phase over all ice surfaces as it was hoped. It is apparent there is a degree of over tting to the Greenland data. Although the Arctic cloud make-up is known to have a large amount of mixed multi-layered clouds (see Shupe, 2002) compared to Antarctica, the drop in performance has a di erent reason, and it seems to be location dependent | see gure 9. The high altitude regions (above 3 km) are challenging for the model, as it was trained only on ice surfaces up to this altitude in Greenland. Training the models on all ice surfaces globally could solve this problem, and so could perhaps including elevation in the predictor set. The high altitude ice sheets are colder and therefore more easily confused with clouds, especially if they are thin and close to said surface. Another way of decreasing the models' bias towards Greenland, at greater computational cost, is to perform K-fold cross-validation for all the intermediary steps (stepwise feature selection and hyperparameter grid search) instead of using a predetermined validation set.

The focus of further work should therefore be the engineering of better predictors that can separate the spectral signatures of the cold surface from that of clouds | especially for thin and low-lying clouds.

Figure 9: The accuracy of the Greenland-trained RF nighttime model is the lowest in the highest altitude regions of Antarctica. A few terrain contour lines are shown, with altitude units in kilometers.

The inclusion of the beta parameters as discussed in Baum et al., 2012 is a priority, as well as the reconstruction of the 1:6 µm Aqua MODIS band with help from Gladkova et al., 2011. Other avenues of further investigation will be more evident if the models are biased towards the Clear class | the majority of misclassi cations will then be of real clouds, whose properties can be studied using CALIOP. Overall, better signal to noise ratios can be obtained by averaging the MODIS spectral values from 1 km resolution to 5 km, not just keeping the middle sub-pixel in the 5 5 grid. It is also worth trying other types of classi ers, such as deep neural networks, which do better when the data is not linearly separable, but are less interpretable.

5 Conclusion

Two RF models were trained to give pixels classi- cations of cloud-mask and cloud-phase over Greenland by using MODIS 5 km spectral observations. A nighttime model using infrared bands between 3:9 µm and 14:2 µm and a daytime model using short-wave infrared (SWIR) and infrared bands between 0:9 µm and 8:5 µm were trained separately. The reference labels for pixels are obtained from collocated CALIOP level 2 5 km merged layer product data for the period of 2013{2017, by applying a strict quality control process. Additional physically relevant predictors were created by processing and combining spectral information from di erent bands, and a subset of the most potent predictors was systematically chosen by using a stepwise feature selection algorithm. The data was carefully split into training, validation and test sets in a way that minimises spatial autocorrelation, and the performance of the models was evaluated on the test set only at the end of the tuning process to produce an unbiased score. The model con guration | its hyperparameters | were optimised by doing a grid search over a range of values. The nal models use 10 predictors each and exhibit great performance over ice surfaces: 83.83% agreement with CALIOP for the RF nighttime model and 87.98% agreement for the RF daytime model, with about 3% of pixels left undetermined by both models. Compare this to the MODIS cloud product algorithms performance of 52.06% for the all-day infrared algorithm and 71.22% for the daylight-only algorithm, which leave about 21% and 4% of pixels undetermined, respectively. In terms of the types of clouds that are frequently missed by the models, both models tend to inaccurately classify geometrically thin clouds close to the surface, whereas the daytime RF model also encounters di culties with high-altitude thin clouds. The models trained on Greenland have incomplete carry-over when tested on Antarctica, with a drop in CALIOP agreement scores: 75.39% for the RF nighttime model and 85.23% for the RF daytime model | still outperforming the MODIS cloud-phase algorithms in the region. More sophisticated ways to systematically select the predictor and hyperparameter can help reduce the discrepancy, but the di erent topographies and weather systems in Antarctica versus the Arctic warrant training a new set of models on all permanent ice surfaces globally. Further work should be focused on feature engineering of physically relevant predictors and reconstructing the information from the missing 1:6 µm Aqua MODIS band. Clues for further investigation can be obtained by biasing the RF models towards correctly classifying clear scenes, and exploring the properties of the misclassi ed clouds using data from CALIOP. Overall, this work proves the great potential of machine learning algorithms for cloud detection and cloud-phase classi cation using passive imager spectral information with the help of more accurate active instruments like CALIOP.

References

Ackerman, S. A., Holz, R. E., Frey, R., Eloranta, E. W., Maddux, B. C., and McGill, M. (2008). Cloud detection with MODIS. part II: Validation. J. Atmos. Ocean. Technol., 25(7):1073{1086.

Baum, B. A., Menzel, W. P., Frey, R. A., Tobin, D. C., Holz, R. E., Ackerman, S. A., Heidinger, A. K., and Yang, P. (2012). MODIS cloud-top property re nements for collection 6. J. Appl. Meteorol. Climatol., 51(6):1145{1163.

Chan, M. A. and Comiso, J. C. (2013). Arctic cloud characteristics as derived from MODIS, CALIPSO, and CloudSat. J. Clim., 26(10):3285{3306.

Frey, R. A., Ackerman, S. A., Liu, Y., Strabala, K. I., Zhang, H., Key, J. R., and Wang, X. (2008). Cloud detection with MODIS. part i: Improvements in the MODIS cloud mask for collection 5. J. Atmos. Ocean. Technol., 25(7):1057{1072.

Gladkova, I., Grossberg, M., Bonev, G., and Shahriar, F. (2011). A multiband statistical restoration of the aqua MODIS 1.6 micron band. In Shen, S. S. and Lewis, P. E., editors, Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XVII. SPIE.

Ji, C. and Ma, S. (1997). Combinations of weak classi ers. IEEE Trans. Neural Netw., 8(1):32{42.

Lee, J., Shi, Y. R., Cai, C., Ciren, P., Wang, J., Gangopadhyay, A., and Zhang, Z. (2021). Machine learning based algorithms for global dust aerosol

detection from satellite images: Inter-comparisons and evaluation. Remote Sens. (Basel), 13(3):456.

Liu, Z., Kar, J., Zeng, S., Tackett, J., Vaughan, M., Avery, M., Pelon, J., Getzewich, B., Lee, K.-P., Magill, B., Omar, A., Lucker, P., Trepte, C., and Winker, D. (2019). Discriminating between clouds and aerosols in the CALIOP version 4.1 data products. Atmos. Meas. Tech., 12(1):703{734.

Maddux, B. C., Ackerman, S. A., and Platnick, S. (2010). Viewing geometry dependencies in MODIS cloud products. J. Atmos. Ocean. Technol., 27(9):1519{1528.

Marchant, B., Platnick, S., Meyer, K., Arnold, G. T., and Riedi, J. (2016). MODIS collection 6 shortwave-derived cloud phase classi cation algorithm and comparisons with CALIOP. Atmos. Meas. Tech., 9(4):1587{1599.

NASA-LaRC (2024). User's guide for CALIPSO. https://www-calipso.larc.nasa.gov/resources/calipso_users_guide/data_summaries/index.php .

Parol, F., Buriez, J. C., Brogniez, G., and Fouquart, Y. (1991). Information content of AVHRR channels 4 and 5 with respect to the e ective radius of cirrus cloud particles. J. Appl. Meteorol., 30(7):973{984.

Platnick, S., Li, J. Y., King, M. D., Gerber, H., and Hobbs, P. V. (2001). A solar re ectance method for retrieving the optical thickness and droplet size of liquid water clouds over snow and ice surfaces. J. Geophys. Res., 106(D14):15185{15199.

Platnick, S., Meyer, K. G., King, M. D., Wind, G., Amarasinghe, N., Marchant, B., Arnold, G. T., Zhang, Z., Hubanks, P. A., Holz, R. E., Yang, P., Ridgway, W. L., and Riedi, J. (2017). The MODIS cloud optical and microphysical products: Collection 6 updates and examples from terra and aqua. IEEE Trans. Geosci. Remote Sens., 55(1):502{525.

Ploton, P., Mortier, F., Rejou-Mechain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., and Pelissier, R. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nat. Commun., 11(1).

Remer, L. A., Kaufman, Y. J., Tanre, D., Mattoo, S., Chu, D. A., Martins, J. V., Li, R.-R., Ichoku, C., Levy, R. C., Kleidman, R. G., Eck, T. F., Vermote, E., and Holben, B. N. (2005). The MODIS aerosol algorithm, products, and validation. J. Atmos. Sci., 62(4):947{973.

Shupe, M. D. (2002). Cloud radiative heating rate forcing using pro les of retrieved arctic cloud microphysics. In Cloud Radiative Heating Rate Forcing Using Pro les of Retrieved Arctic Cloud Microphysics, St. Petersburg, Florida. Atmospheric Radiation Measurement (ARM), Atmospheric Radiation Measurement (ARM).

Vaughan, M. A., Powell, K. A., Winker, D. M., Hostetler, C. A., Kuehn, R. E., Hunt, W. H., Getzewich, B. J., Young, S. A., Liu, Z., and McGill, M. J. (2009). Fully automated detection of cloud and aerosol layers in the CALIPSO lidar measurements. J. Atmos. Ocean. Technol., 26(10):2034{2050.

Wang, C., Platnick, S., Meyer, K., Zhang, Z., and Zhou, Y. (2020). A machine-learning-based cloud detection and thermodynamic-phase classi cation algorithm using passive spectral observations. Atmos. Meas. Tech, 13(5):2257{2277.

Winker, D. M., Vaughan, M. A., Omar, A., Hu, Y., Powell, K. A., Liu, Z., Hunt, W. H., and Young, S. A. (2009). Overview of the CALIPSO mission and CALIOP data processing algorithms. J. Atmos. Ocean. Technol., 26(11):2310{2323.

Appendices

A Detailed quality control process

The first step is to eliminate profiles contaminated with aerosols, as it is known to interfere with cloud retrieval in passive sensing applications. As such, only aerosol-free profiles are kept, defined as profiles having a column aerosol optical depth (AOD) smaller than 0.05. Profiles 2, 3, and 7 in figure 3 would be eliminated, for example.

Each layer is also given a cloud-aerosol discrimination (CAD) score by a feature detection algorithm described in detail in Liu et al., 2019. Only profiles containing layers with a CAD score larger than 70 out of 100 are kept, which signify high-confidence cloud detection.

Each layer classified as cloudy is further given a phase classification (Water, Ice, Oriented ice and Unknown). With this classification comes a set of quality assurance flags, from 0 (no/low confidence) to 3 (high confidence) for cloud-phase. Only profiles containing relevant layers with the highest quality assurance are kept.

Aerosol-free profiles are classified as *Clear* only if they contain no detected feature layers | only profile 6 in figure 3 would be classified as *Clear*. Aerosol-free profiles are classified as *Water* if they only contain cloud layers with CAD > 70 with a water phase classification of the highest quality (quality assurance flag has a value of 3). The same is done for the *Ice* classification | profiles 1 and 4 would qualify. However, there are a significant amount of profiles (about 10-15% over Greenland) that contain some layers classified as water clouds, and some layers classified as ice clouds: these are temporarily saved as *Mixed multilayer* classification | these would be profiles 5 and 8.

Testing has shown that the ML model with MODIS data as input is poor at distinguishing between the *Mixed multilayer* type and the Water and Ice types, whereas *Water* and *Ice* types are reliably predicted. This is not surprising, passive imagers struggle to extract vertical information, even using clever techniques like CO_2 slicing and Infrared Window, as explained in Baum et al., 2012. Instead of discarding a significant number of profiles classified as *Mixed multilayer*, a decision process was created to reclassify some of them as *Water* or *Ice*. Any *Mixed multilayer* profile will be determined to be water-dominant or ice-dominant by using their respective layers' integrated attenuated backscatter (IAB) coefficient as a proxy for quantity of water or ice content in the profile. More specifically, if the ratio of the total IAB in the ice layers to the total IAB in the water layers is greater than a manually chosen threshold of 5, then the profile is reclassified as ice. If the inverse of this ratio is larger than 5, then the profile is reclassified as water. If none of these conditions are met, the cloud-phase is considered ambiguous and the profile is discarded.

for every *Mixed multilayer* profile:

$$\text{ratio} = \frac{\sum_{\text{ice}} IAB_{\text{layer}}}{\sum_{\text{water}} IAB_{\text{layer}}}$$

if ratio > 5, profile is ice

if ratio < 1/5, profile is water

else, discard profile

With this method, about half of the *Mixed multilayer* profiles are recovered.

B Random Forests and decision trees

Decision trees can be fitted to data, but they have two large drawbacks: they are not as accurate as other classification methods (due to the greedy binary space partitioning approach) and they tend to overfit to the training data, meaning that completely different decision trees can be created from two slightly different training datasets. A natural proposal to solve this problem is then to train many decision trees, each on a slightly different version of the training set obtained by bootstrapping - a resampling method with replacement. This is called bagging, from bootstrap aggregating. However, this is not enough to reduce the variance of models with respect to the training data, as many trees will still be highly correlated, with the strongest predictors always closer to the root of the tree. The Random Forest solves this problem by also restricting the number of features accessible to each decision node in the training process. This way, the trees become decorrelated, and the model is much less prone to overfitting.

>

Figure 10: First four levels of a decision tree in the RF nighttime model. The split decision is shown at each node, along with the distribution of datapoints in the node variable for each class.

C Stepwise Feature Selector

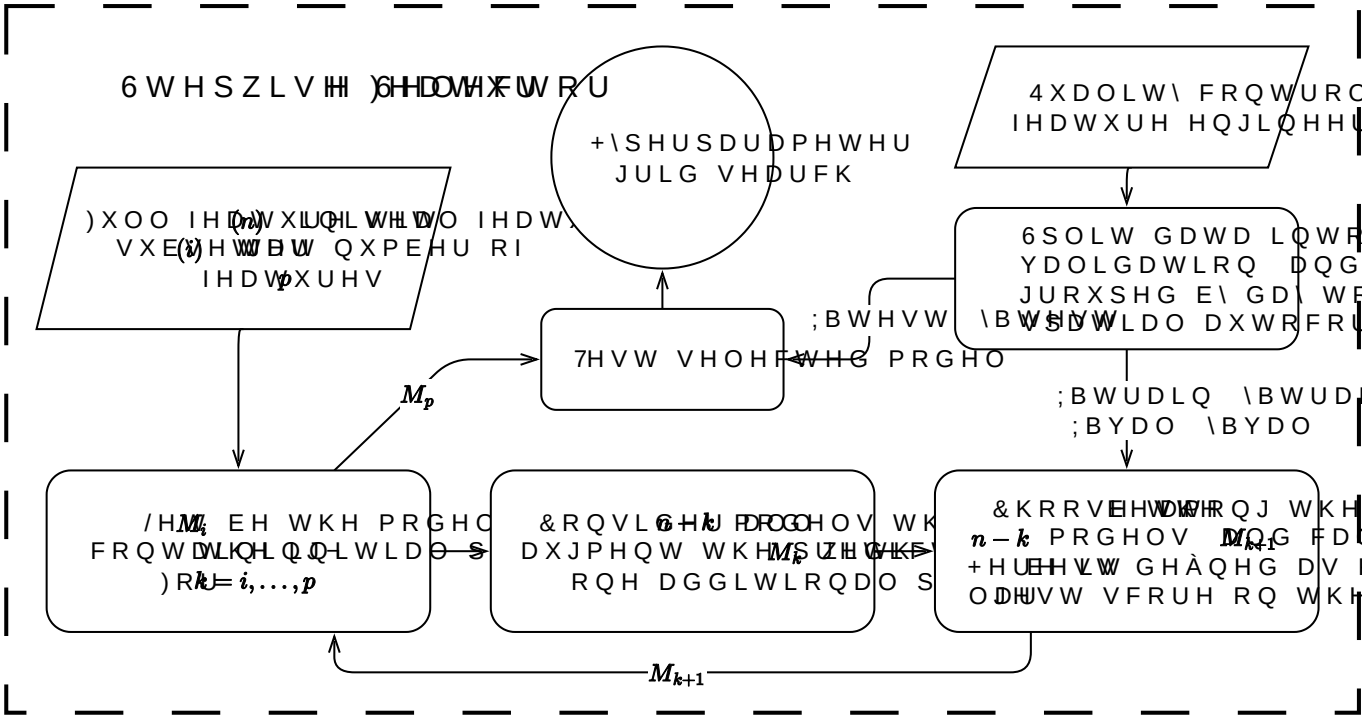

Figure 11: A diagram explaining the Stepwise Feature Selection algorithm.

D Hyperparameter grid search

The models with the ten most relevant predictors have been trained with a range of different hyperparameters, as shown in figure 12.

N_{depth}, with values from 10 to unlimited, number of trees in the forest N_{trees}, with values from 50 to 500, and the number of minimum samples at a leaf node N_{leaf}.

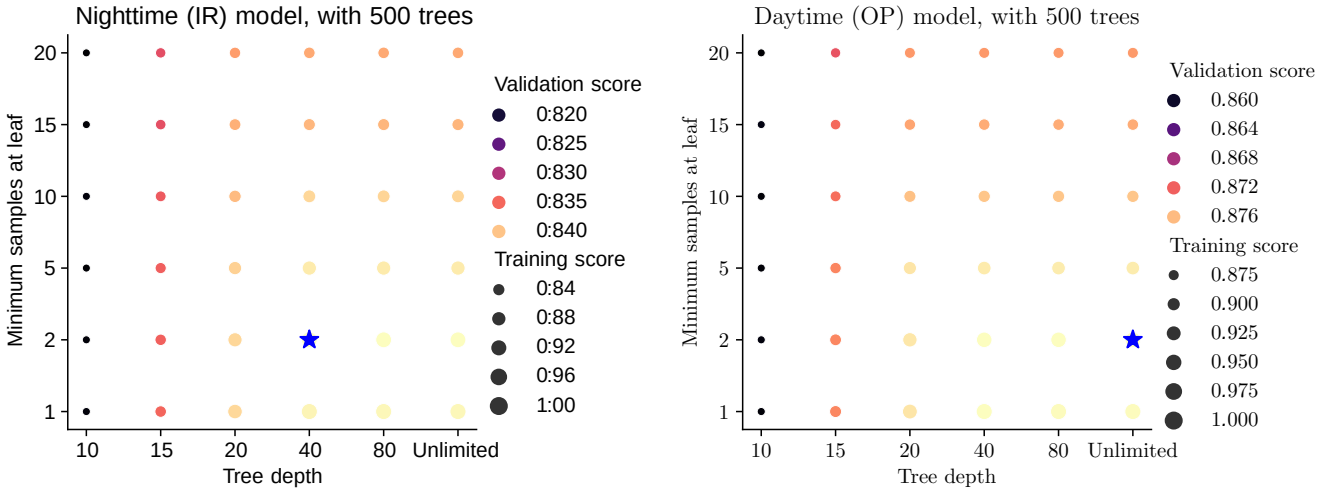

Figure 12: The validation (in colour) and training (by size) scores as a function of N_{depth} and N_{leaf}. A variation of N_{trees} was done separately to reduce computation time, as the number of trees, above a certain amount, will not improve or hurt the model performance significantly.